Using machine learning to predict if a customer's account is default

By Jack Skelton

# Tables of contents

[Tables of contents 1](#_Toc191300504)

[Project aim 2](#_Toc191300505)

[Project expectations 2](#_Toc191300506)

[Literature Review 5](#_Toc191300507)

[Methodology 12](#_Toc191300508)

[Methodologies researched 12](#_Toc191300509)

[Methodologies comparison 14](#_Toc191300510)

[Software researched 14](#_Toc191300511)

[Software comparison 14](#_Toc191300512)

[Machine Learning models chosen 14](#_Toc191300513)

[Methodology chosen/Framework 14](#_Toc191300514)

[Data 14](#_Toc191300515)

[Results 14](#_Toc191300516)

[Conclusion 15](#_Toc191300517)

[Appendix 15](#_Toc191300518)

[Risk register 15](#_Toc191300519)

[Bibliography 15](#_Toc191300520)

# Project aim

My project aim to develop an AI using the software R studio to make predictions on a customer account to see if it is default. This project would allow banking organizations to predict which accounts are at risk of becoming default from missing payments, this will be done to allow the banks to minimize the losses they would have received without the predictions. I will be developing my R software script skills throughout this project learning on how to implement different machine learning modules into python scripts by creating the code for them. This would be an appropriate challenge for level 6 since I will be using machine learning which may not have complex code but will have great understanding on what happens in the process which will make my understanding better for my studies since I will be studying machine learning in semester 2.

# Project expectations

Investigation

In this investigation I will be creating 5 different R scripts which will include 5 different machine learning modules. These will include these individual modules Naive Bayes, Decision Tree, Random Forest, Regression, KNN. I will then take a data set from archive.ics.uci.edu to get some banking details. I will then use this data set and put it into the R studio script to teach the AI to predict if an banking account will be at the risk of becoming default. I will do this by looking at the dataset for banking details to look at the past payment and other data to make predictions to see if a user is going to miss a payment to allows the banks to find out what accounts are at risk. I will then compare the different machine learning modules predictions to identify what machine learning modules makes the best predictions and this python script will be my final product. I will need to do a video at the end of the project to display how the code on the different R scripts works and how the predictions are made, I will also have to show how I identified which machine learning module makes the best predictions. The data set I will be using will be Default of Credit Card Clients (see appendix B for link or click data set name).

Evidence good practice

I will evidence my findings by using a report documenting all my findings on the machine learning predictions and I will also be making a video to record my findings on what the predictions where and what the best machine learning module was to predict the risk of an account becoming default. I will also produce a monthly project report to document what has happened every month during the production module, I will also produce a use of generative AI which will document the use of AI like ChatGPT throughout the production module which will show what queries I have asked an AI to get some answers. All this would show good practice and allow the project to be successful. I will also have to stay within the laws like the computer misuses act since I will be using a dataset from someone else so the data is not mine, thus I will need to be careful about what I use the data for to not breach any laws or morals.

Appropriate evaluation

An appropriate evaluation of this project can be broken down into several aspects: relevance, importance, cost and quality. My project is relevant to modern times since I will be using and creating some machine learning models that is a part of an industry that is growing and most things now have some connections to AI, so this would be good to do since I will help me find a career after university with important skills. My project will only cost me time, since no money is needed to complete my project since the only bit of software I am using is free, though this project will cost me time since I will need to spend a lot of time researching the different machine learning models and creating and testing the scripts. The quality of the project is up to me and is impacted by how much time is spend on the project, having a good quality product would mean that the project was a success and that I could then use the skills developed to help me in the future and would be a nice talking point in an interview. The importance of this project is quite high since if this is a successful project then the project produced out of the project would be helpful to the banks and would allow them to reduce their losses and make sure they could update any users for the risk that they are approaching with their account so they done become bankrupt and have no outstanding debts.

A MoSCoW type requirements analysis or equivalent

|  |  |  |  |
| --- | --- | --- | --- |
| **MUST**  (40-50%) | **SHOULD**  (50-70%) | **COULD**  (70%+) | **WON’T**  (To provide clarity on what won’t be included) |
| * Create 5 different machine learning models * Test ai for predictions * Evaluate which ai is best for predction * Document process | * Create ai models in R script * Test ai for crime rate prediction * Only submit one machine learning model for my product | * Have hybrid machine learning models to predict crime rate * Use data other than classification | * More than 5 machine learning models * Using different programming languages |

# Literature Review

|  |  |  |  |
| --- | --- | --- | --- |
| **Research Objective** | **Methodology/Tools Used** | **Results** | **Paper Reference Number** |
| Predict TBM operation parameters using machine learning | Decision tree | Proposed a BAS algorithm  Improved rationality  Reduced uncertainty  Better predictions | Wang, Y. et al. |
| Machine Learning Models for Predicting the Ammonium  Concentration in Alluvial Groundwaters | Classification  Regression  Baseline models | Deep Neural Network outperformed other model in predictions, this was shown with the plots comparison | Perović Marija et al. |
| Prediction of metastatic pheochromocytoma and paraganglioma | Decision tree | Highest performance is achieved by SVM and LogReg, SVR and RR models | López Steinmetz, L.C. et al. |
| Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine | Dual methodology | Shows pre existing depression and anxiety  Machine learning models show potential at risk students | Pamporaki, C. et al. |
| Open AccessEditor’s ChoiceReview  Flood Prediction Using Machine Learning Models: Literature Review | Flowcharts | Loads of articles reviews and studies were analyses  There was an effective use of machine learning models | Mosavi, A., Ozturk, P. and Chau, K.-w. ( |
| Designing Disease Prediction Model Using Machine Learning Approach | K-Nearest neighbor, Convolutional neural network | The system shown in this method worked on showing the disease risk prediction  Machine learning models have a good comparison showing accuracy and processing time of both models  the system shown has the potential to show large scale medical data effectively | Dahiwade, D., Patle, G. and Meshram, E. |
| Disease Prediction using Machine Learning Algorithms | Disease Tree Classifier  Random forest Classifier  Naïve Bayes Classifier | This article shows how accurate the algorithms is at 95% using medical classification  This model has some future proof since the more medical data releases the more accurate the algorithm will be | Grampurohit, S. and Sagarnal, C. |

|  |  |  |
| --- | --- | --- |
| **Previous work** | **Limitations** | **Proposed System**  **(How it will cover those limitations** |
| Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine | Further studies are needed before clinical implementation | Make sure that the data set is proper and make sure there are no empty data fields |
| Open AccessEditor’s ChoiceReview  Flood Prediction Using Machine Learning Models: Literature Review | Machine learning was in early development  There are some data limitations in this method  Some machine learning models require good optimization for it to work | Machine learning is still growing but in a better place now  Check dataset for any data limitations  Make sure I am using a suitable software for the coding |
| Designing Disease Prediction Model Using Machine Learning Approach | They only used two different machine learning models  The system proposed depends on the data sets quality and size | I will be using multiple different machine learning models  Make the product able to handle different data sets |
| Disease Prediction using Machine Learning Algorithms | This study only uses 3 models limiting the potential performance  The model relies on quality and quantity of the data  Matrix are useful but do not fully capture model effectiveness | Will be using more models to predict  Make the product handle more data sets  Look at other metrics like precision or recall |

**A screenshot of a computer

Description automatically generated**

**This photo shows the results for the Disease Prediction using machine learning**

|  |  |  |
| --- | --- | --- |
| **Paper types** | **Methodology** | **Year** |
| Article | Decision tree | 2023 |
| Article | Classification  Regression  Baseline models | 2021 |
| Article | Decision tree | 2023 |
| Article | Dual methodology | 2024 |
| Article | Flowchart | 2018 |
| Article | K-Nearest neighbor  Convolutional neural network | 2019 |
| Thesis, Dissertation | Disease Tree Classifier  Random forest Classifier  Naïve Bayes Classifier | 2020 |

One of the papers I have had a look at was a paper using the same dataset and a similar project idea, so this paper can be a helpful tool to look at how a past project doing this has taken a approach which I can then look at and see if I can make it better, a key thing I will need to avoid is that I need to make sure my project is not just copying this paper. The paper I am talking about is called “Credit Default Mining Using Combined Machine Learning and Heuristic Approach” this paper was published by Sheikh Rabiul Islam and W. Eberle, S. Ghafoor 2018. The aim for this project was to look at a dataset and predict if an account is default. A default account is an account made by a user that has missed some payments and is closed by the bank. This is helpful since it allows the banks to see what accounts could be at risk of becoming a default account and the probability of a payment in advanced. The paper than states that the reason this is done is to “earlier the potential default accounts are detected the lower the losses” this would then show why it is helpful for banks to make these predictions since it will make the losses through these accounts lower so they don’t lose as much money. Another thing this paper states about the reason for this project is so that it “could investigate and help the customer by providing necessary suggestions to avoid bankruptcy and minimize the loss” this clearly shows that the project is used to help the customer and the bank on both ends since as discussed before it helps the bank lower the losses and shows the customer some helpful things that can be done to help with their payments.

When it comes to this papers methodology they implement 2 different ones. This is a machine learning approach and a heuristic approach. These are then discussed on what these different methodologies are and how they could be used. That machine learning approach is said to be used to compare different machine learning models and see which one is most effective at making their predications, the other one is used to make the predictions since it is used throughout the different machine learning models to make the predictions of what accounts are at risk of becoming a default accounts. This is helpful for my project since it allows me to look a some methodologies and see which would be suitable for my project, these would be helpful since I will be using multiple different machine learning models to compare the different effectiveness of the models to create my final product, and the other one could be used to create the predictions that I will be making in the future.

The results of this paper shows that both methods used in this project were successful at making there predictions. The results shown from this project show that the methods used outperform currently used methods. This shows that there project was a great success and that banking organisations should implement their ideas.

A graph with colorful bars

Description automatically generated with medium confidence

Image taken from paper “Credit Default Mining Using Combined Machine Learning and Heuristic Approach”

# Methodology

## Methodologies researched

### Machine Learning Approach

Looking at one of the previous papers credit default mining using combined machine learning and heuristic approach they have an approach which was used to determine what machine learning model used was the best for the predications they were trying to make but while looking at the data shown in the paper there was no steps shown what was taken to determine this. So I did some extra research for machine learning approaches and had a look at the steps taken in the method. I found a webpage showing an approach for this method as seen in the image below. This method contains 5 major steps:

1. Collection of data
2. Data cleaning & feature engineering
3. Model building
4. Evaluate
5. Model deployment

Breaking these steps down we can see that in the first step that this is where the user is collecting the data which for me was finding the data set. Step 2 is to look at the data set for any errors in the data like null values. Step 3 is the step done to program the models used for predictions. Step 4 is looking at the results of the each model to determine what was best at making predictions and step 5 is the step where the product is deployed into a working environment.

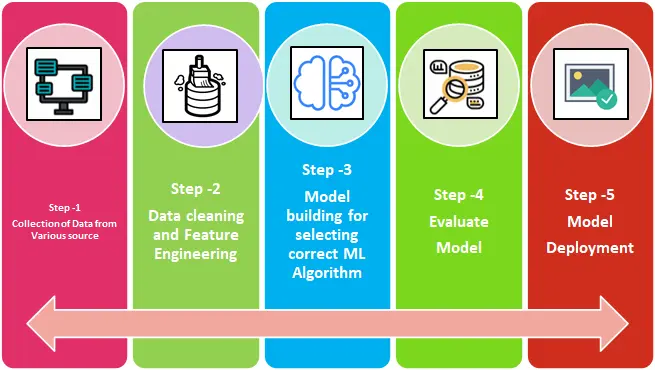


Image of methodology steps taken from labellerr.com

Labellerr. (2022). *Machine learning steps: A complete guide for beginner in ML*. [online] Available at: https://www.labellerr.com/blog/machine-learning-steps-a-complete-guide-to-the-ml-process/.

### Heuristic Approach

The paper looking at prediction of default credit also had a heuristic approach, which is a method that contains two tests a standard test and a customer specific test, these are used to find out which accounts were at the potential risk of becoming a default account. This model was used for two steps one to determine the potential risk of a account becoming a default account by looking at the transitional history. And the other step was to look at the real time transitions of an account to determine a score of the account becoming a default account. Though this paper did not include a diagram of this model when we look at the heuristic approach online we can see what the models looks like in the image below, as we can see that the model has 4 steps to it. The first one is used to look at the problem and understand it by creating a scope, the second one is making a plan to mitigate the problem, step 3 is to carry out the plan created, and the last step is used to evaluate and adjust the method used.

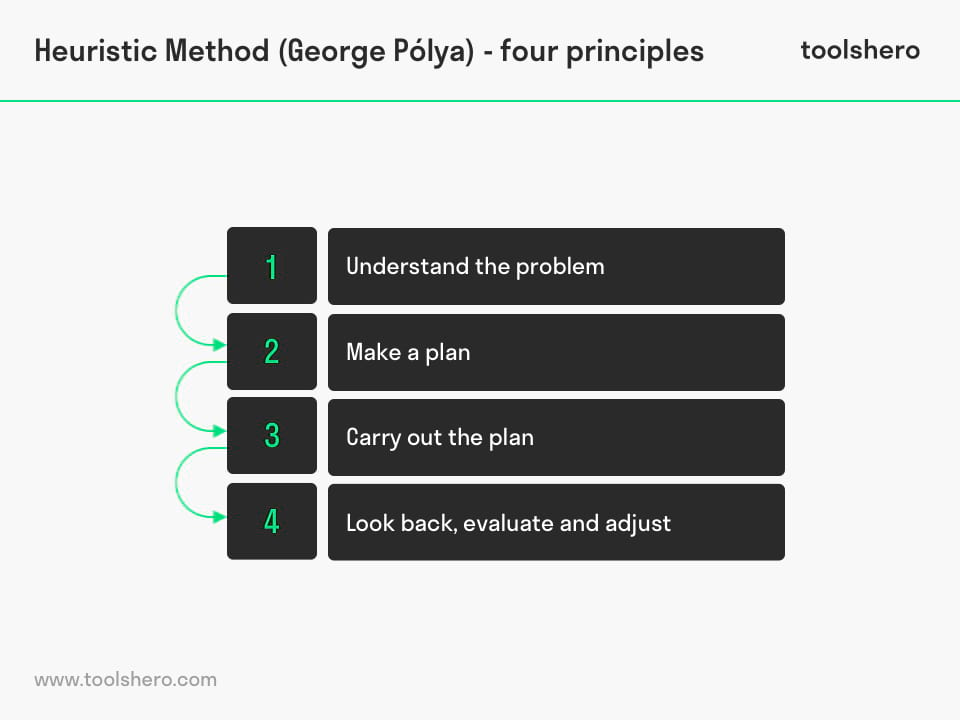


Image taken from toolshero.com

Mulder, P. (2018). *Heuristic Method, a problem-solving method | ToolsHero*. [online] toolshero. Available at: https://www.toolshero.com/problem-solving/heuristic-method/.

### Conceptual System

This method was used in a paper to determine the crop yield based on machine learning models. This model was not discussed in the paper only showing a diagram of how the method was done, this can be seen in the image below. As we can see that the data sourced to the user was then made into a centralized dataset via ETL. Then the models were made and trained using the dataset, the models were also evaluated at this stage of the method to determine what was the best prediction-based model. This was then implemented into a predictions system where it will be feed new data to make predictions based on crop yield.

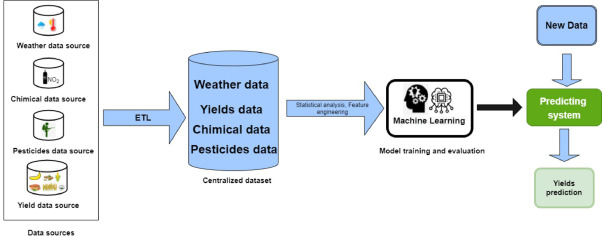


Image taken from sciencedirect.com crop yield prediction based on machine learning models.

## Methodologies comparison

Looking at these methods we can see that they all include machine learning models to make predictions based on a data set and each had an evaluation step to determine what the best model for predictions was for the product. Though there are some differences with the heuristic approach, it is the only one that includes a step which has the user go back and adjust the method used to make the predictions better. The conceptual system was the only method used which had at the end of the model training feeding new data to make predictions and is the only method that has been shown to use ETL in the dataset, though this could be included in the second step of the machine learning approach where it is said that the user must clean the data.

## Software researched

These are some of the software that I researched that I could use to complete this project.

|  |  |  |
| --- | --- | --- |
|  | R studio | Python |
| Areas of use | Popular in academic and businesses | Well suited for many types of programs |
| Research, finance and data science | Data science, web development, software development and gaming |
| Language | General-purpose programing language | General-purpose programing language |
| Advantages | * Open source * 19,000 packages * Easier to learn at start * Organized interface * Great for graph making * Many functionalities for data analysis * Great for statistical analysis * Using in machine learning module | * Open source * +300,000 packages * Beginner friendly * English like syntax * High ease of deployment and reproducibility |
| Disadvantages | * More difficult when using advanced functionalities * Slower language * Finding right libraries can be hard | * Poor memory efficiently * Has less libraries for data science * Real time errors * Needs rigorous testing * Visualizations are more convoluted |

Comparing the 2 software there are some similarities with them like that there are both general purpose programming languages. But R studio is made for research, finance and data science, which is was I care about in the software for data science since this is was my predictions models are in. python is still a good software to know but for this project the better functionalities for data science like graphs and analysis make this a no brainer in which software to pick for this project.

## Machine Learning models chosen

Chosen models:

1. Naïve Bayes
2. Decision Tree
3. Random Forest
4. Regression
5. K Nearest Neighbors (KNN)

## Methodology chosen/Framework

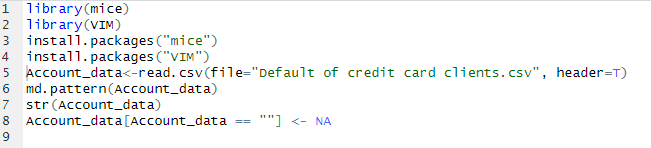
When it comes to my methodology that I am going to use for my project, I am going to user the machine learning approach which will include these steps.

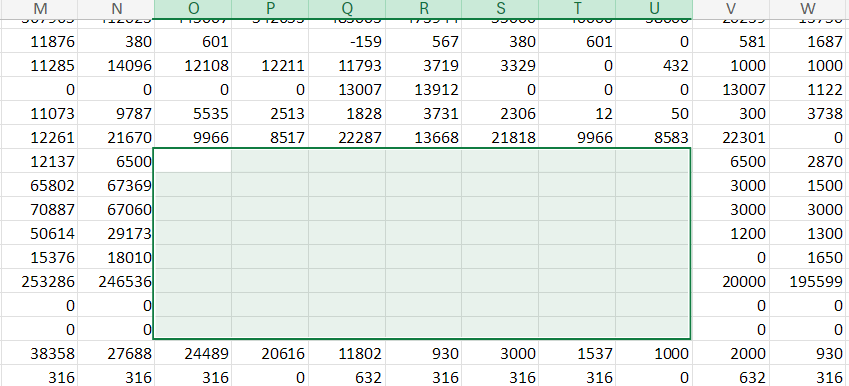
1. Data collection
2. Data cleaning
3. Model building
4. Evaluate
5. Model deployment

These steps will be used to complete this project where I will use the data set I have chosen to complete the data collection step of this model, I will then open the dataset in R studio to then check if the data has any null data contained in the dataset and use methods to replace the data so that there are no errors, in the next step I will then write the code for each of my chosen models and start running them to make predictions on which accounts are at risk of becoming default. I will then use code in R studios to make graphs to show where there could be errors in the data made from the predictions to make a good evaluation of the models used to determine which model is going to be my product. Then for the final step I will upload my code on to git hub through my process to make sure that there is come records of time and that my final product is outlined for other users to take and deploy onto their systems, in these last couple steps I will need to make a presentation about my product which will include my evaluation of each model and a dashboard to make it presentable.

# Data

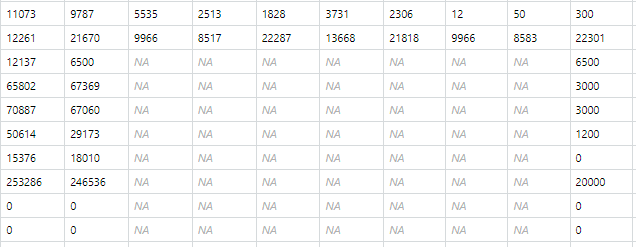
In this section I will look at the dataset taken for my project and look though it for missing data to then fill the spots in via KNN. This is done to make the predictions more accurate since there won't be any null data in the dataset

This is the code I used to allow R studio to gain access to the data set and check for any null data. This was done by downloading two libraries these are mice and VIM, the library mice is used to use the md.pattern function to check the data set for any null values, the second library is VIM this is used for KNN predictions to fill these missing values. Line 5 was used to allow R to read the data set making it a variable.

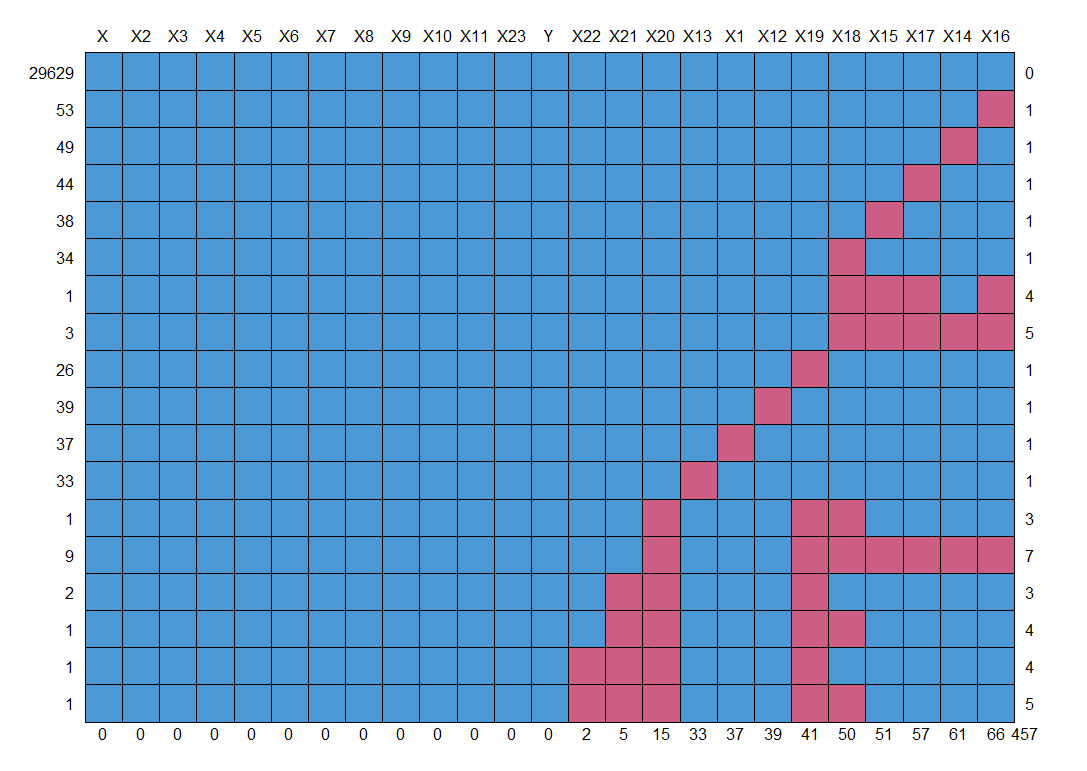
Looking at the dataset there is some missing values but they are not in the correct format since we will need them to be N/A to make them show up while using the md.pattern function. To fix this issue in the data set we will need to change this data to a correct format, this was done by this line of code:



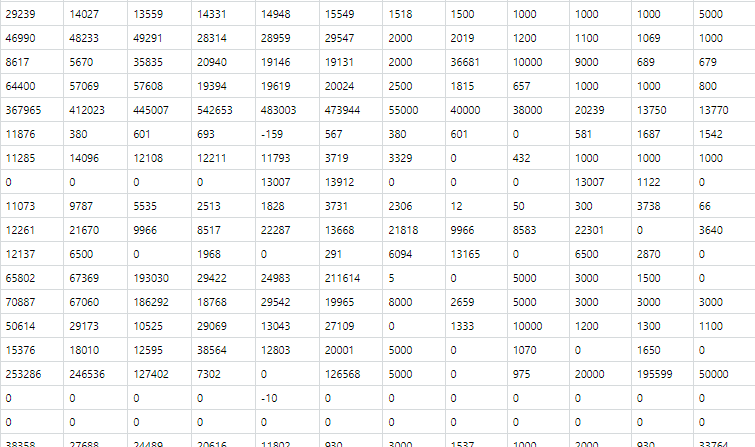
After doing this we can see that the data has been changed.

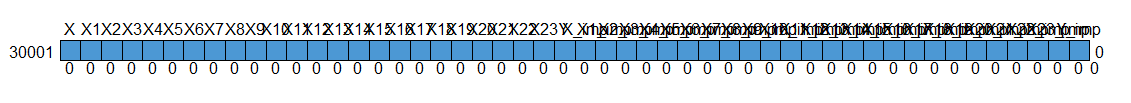
After this we can use md.pattern to see where the data is missing in the data set.



Now we can start filling in these null data.

Line 9 is the code that is used to fill in the null data using KNN, I had to create a new variable to which is the new imputed data created. After this we can then look at the summary of the data to see if the data has been changed:

But the best way that this can be easily seen is by using md.pattern to make a graph to see any null data:

now after this we will need to write this new imputed data into the data set for us to call back to on a later date. This was done by the code below which created a new file called Default of credit card clients cleaned.csv:



# Results

# Conclusion

# Appendix

## Risk register

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Risk Description** | **Likelihood** | **Impact** | **Owner** | **Mitigation** | **Status** |
| **1** | **Loss of data** | Low | High | **Jack** | **Save data to cloud** | **Open** |
| **2** | **Software accessibility** | Low | High | **Jack** | **Download software** | **Closed** |
| **3** | **Miss use of data** | Medium | High | **Jack** | **Learn about legistrations and how to follow them** | **Open** |
| **4** | Dataset could be full of errors | Medium | High | Dataset owner/ Jack | Inspect data set to see if there are any errors and send feedback to the owner to make appropriate changes | Open |
| **5** | Product Implementation | Medium | High | Police Force | The Banks would need to check if they could implement my product into their system | OPEN |

## Bibliography

Data set: Default of Credit Card Clients

Yeh, I. (2009). Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C55S3H>.

Wang, Y. et al. (2023) “Prediction of TBM operation parameters using machine learning models based on BPSO,” Advanced Engineering Informatics, 56. Available at: <https://doi.org/10.1016/j.aei.2023.101955>.

Perović Marija et al. (no date) “Machine Learning Models for Predicting the Ammonium Concentration in Alluvial Groundwaters,” Environmental Modeling & Assessment, 26(2), pp. 187–203. Available at: <https://doi.org/10.1007/s10666-020-09731-9>.

López Steinmetz, L.C. et al. (2024) “Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine,” Frontiers in Psychiatry, 15. Available at: <https://doi.org/10.3389/fpsyt.2024.1376784>.

Pamporaki, C. et al. (2023) “Machine learning models for prediction of metastatic pheochromocytoma and paraganglioma: a cross-sectional cohort study,” The Lancet. Digital health, 5(9), pp. e551–e559. Available at: <https://doi.org/10.1016/S2589-7500(23)00094-8>.

Mosavi, A., Ozturk, P. and Chau, K.-w. (2018) ‘Flood prediction using machine learning models: Literature review’, *Water*, 10(11), p. 1536. Available at: <https://doi.org/10.3390/w10111536> (Accessed: 03/02/25).

Dahiwade, D., Patle, G. and Meshram, E. (2019) ‘Designing disease prediction model using machine learning approach’, *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 1211–1215. Available at: https://doi.org/10.1109/ICCMC.2019.8819782 (Accessed: 03/02/25).

Grampurohit, S. and Sagarnal, C. (2020) ‘Disease prediction using machine learning algorithms’, 2020 International Conference for Emerging Technology (INCET), Belgaum, India, pp. 1–7. Available at: https://doi.org/10.1109/INCET49848.2020.9154130 (Accessed: 03/02/25).

Islam, S.R., Eberle, W. and Ghafoor, S.K. (2018) 'Credit default mining using combined machine learning and heuristic approach', *ArXiv*, abs/1807.01176.

Labellerr. (2022). *Machine learning steps: A complete guide for beginner in ML*. [online] Available at: <https://www.labellerr.com/blog/machine-learning-steps-a-complete-guide-to-the-ml-process/>.

Mulder, P. (2018). *Heuristic Method, a problem-solving method | ToolsHero*. [online] toolshero. Available at: <https://www.toolshero.com/problem-solving/heuristic-method/>.

Lontsi Saadio, C., Adoni, W.Y.H., Aworka, R., Zoueu, J.T., Kalala Mutombo, F. and Kimpolo, C.L.M. (2022). Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.4003105.

Luna, J.C. (2020). *Python vs R for Data Science: Which Should You Learn?* [online] Datacamp.com. Available at: https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference?dc\_referrer=https%3A%2F%2Fwww.google.com%2F [Accessed 24 Feb. 2025].